MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2018/2019

TDS3551 – DATA MANAGEMENT

(All Sections / Groups)

2 MARCH 2019
2:30 p.m. – 4:30 p.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 7 pages (including cover page) with 5 Questions only.

2. Attempt **ALL** questions. The distribution of the marks for each question is given. This paper carries **100 marks.**

3. Please print all your answers in the Answer Booklet provided in this examination.

Question 1 (20 Marks)

a) Relational Database Management Systems (RDBMS) may have timeouts when data is arriving with high velocity and volume. In order to address this, several techniques are employed by the database engineers. Other than batch queueing the entries, where instead of executing a write for each database entry, the entries can be grouped (batched) together (e.g. 50 entries) and a single write can then be executed for it. This will be more efficient and will require less resources from the RDBMS. What are the other 3 (THREE) techniques? Provide pictorial illustration and describe the techniques.

[12 marks]

[Pictorial illustration (2 marks) and description for each technique (2 marks)]

b) Provide one (ONE) example with an appropriate description for each of the following sources of data.

i. Databases
ii. Files
iii. Social Media
iv. IoT devices

[4 marks]

c) State and briefly explain any 2 (TWO) of the 4 (FOUR) benefits of Open Data Initiative.

[4 marks]

Question 2 (20 Marks)

a) Columnar Based Databases.

i. Differentiate a Columnar Based Database from a regular RDBMS and state an advantage of using a Columnar Based Database.

[2 marks]

ii. State 2 (TWO) Columnar Based Databases products.

[1 mark]

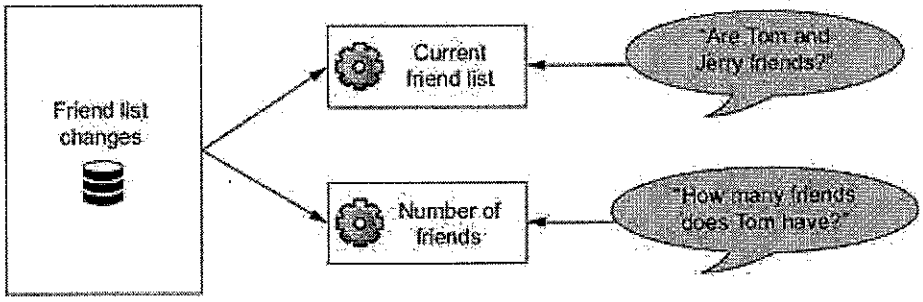
Continued ...

b) Illustrate, using a diagram, the inter-relationship between the Batch Layer, Speed Layer and the Serving Layer. Explain the flow of the incoming and historical data to the queries made by applications.

[6 marks]

c) The figure illustrates the one-way process from Data to Views to Queries. Explain what are Data, Views and Queries.

[3 marks]



d) Using the data table below, develop a fact-based model representation for the following data using exactly 3 (THREE) facts (in other words, 3 fact tables). Explain clearly the rationale behind your model.

User information						
id	name	age	gender	employer	location	timestamp
1	Alice	25	female	Apple	Atlanta, GA	2012/03/29 08:12:24
2	Bob	36	male	SAS	Chicago, IL	2012/04/12 14:47:51
3	Tom	28	male	Google	San Francisco, CA	2012/04/04 15:31:24
4	Charlie	25	male	Microsoft	Washington, DC	2012/04/09 11:52:30

[8 marks]

Continued ...

Question 3 (20 Marks)

a) For the GraphML code below, illustrate it graphically (in other words, draw the graph).

```
<graphml>
 <graph id="grp" edgedefault="directed">
  <node id="1"/>
  <node id="2"/>
  <node id="3"/>
  <node id="4"/>
  <node id="5"/>
  <node id="6"/>
  <edge id="e1" directed="true" source="1" target="2"/>
  <edge id="e2" directed="true" source="2" target="3"/>
  <edge id="e3" source="2" target="4"/>
  <edge id="e4" source="3" target="1"/>
  <edge id="e6" source="5" target="6"/>
  <edge id="e5" source="6" target="3"/>
  <edge id="e7" directed="false" source="4" target="5"/>
  <edge id="e8" directed="true" source="4" target="6"/>
 </graph>
</graphml>
```

[6 marks]

b) Apache Hadoop consists of the Hadoop Filesystem (HDFS) and Hadoop MapReduce. The Hadoop MapReduce is a YARN based system for parallel processing of large datasets.

i. In the high-level organization of HDFS, it consists of nodes that are called DataNodes and NameNodes. Explain the differences between them.

[2 marks]

ii. What does the acronym YARN stands for and what are the functions of YARN.

[2 marks]

iii. What is meant by MapReduce?

[2 marks]

Continued ...

c) The Speed Layer uses incremental algorithms while the Batch Layer uses re-computation as their computation methods. State 3 (THREE) trade-offs with comparisons between the two methods (incremental algorithms and re-computation).

[6 marks]

d) State and explain 2 (TWO) disadvantages of a fully incremental system as oppose to the re-computational system of the Big Data architecture?

[2 marks]

Question 4 (20 Marks)

a) In a Hadoop Distributed File System (HDFS):

i. What is the default data block size in Apache HDFS?

[1 mark]

ii. Each data block is replicated 3 (THREE) times by default. Assuming that there are 4 data blocks represented as {A, B, C, D} and 6 DataNodes, use a diagram to show an example of how the blocks are distributed across the DataNodes (using the default number of replication).

[3 marks]

b) In a Relational Database Management System, as well as when storing data in the batch layer, data normalization is conducted.

i. Explain what is the purpose of data normalization, and

ii. Why would it be better to store the data in de-normalized form in the Serving Layer of the Big Data Architecture?

[4 marks]

Continued ...

c) Batch view is striking a balance between the size of the precomputed views and the amount of on-the-fly computation required at query time. Assuming that data is available for daily sales of a retail store. In precomputing the views, it is decided that in addition to daily, monthly and quarterly sales views will also be precomputed.

i) How many buckets of data is required for the 3 (THREE) granularities (daily, quarterly and yearly) if we are to store 2 years (assuming that it is not a leap year) of the data?

[2 marks]

ii) What is the percentage of additional space (effort) required for creating just the quarterly and yearly buckets, as opposed to just storing the daily buckets?

[2 marks]

iii) Assuming that you would like to compute for a specific date range from 28 March to 5 July (inclusive of both days), which is a total of 100 days (4 in March, 30 in April, 31 in May, 30 in June and 5 in July). If you query from the daily bucket, you would need to access 100 buckets. Explain in detail how you can reduce the number of specific bucket queries using the quarterly and/or yearly buckets.

[6 marks]

iv) For part (iii) above, what is the percentage savings as opposed to accessing it with only the daily buckets stored?

[2 marks]

Question 5 (20 Marks)

a) In the Big Data Architecture, what is the purpose of the Speed Layer?

[2 marks]

b) Which of the desired properties of a Big Data Architecture is specifically addressed by the Speed Layer?

[1 marks]

c) State and give a brief description of the 4 (FOUR) requirements of a Serving Layer.

[8 marks]

Continued ...

d) The Speed Layer stores data for a few minutes or hours. These data are transient data in that they will be removed after a certain period of time. The Speed Layer produces Real-Time views that represent recent data and is continuously updated at a rapid pace.

Assuming that the overall system receives about 2.0 TeraBytes of data a day where data arrives in an average size of 1 MegaByte (assume that it is at a constant pace).

The Batch Layer is scheduled to generate Batch Views 4 times a day.

i. For every Batch View computation, how many times does the Speed Layer needs to do the computation (assuming re-computation method)?

[3 marks]

ii. On average, what is the size of each of the computation?

[2 marks]

e) What does CAP (in CAP Theorem) stands for and briefly explain the CAP Theorem.

[2 marks]

f) In stream processing, differentiate between One-At-A-Time and Micro-Batching processing.

[2 marks]

End of Page